5-15-2014

The Digital Dark Ages: Preserving History in the Era of Electronic Records

Brent West
University of Illinois, bmwest@illinois.edu

The Digital Dark Ages:

Preserving History in the Era of Electronic Records

Brent West
University of Illinois Springfield
bmwest@uis.edu

ABSTRACT

There is mounting concern that, due to the rapid pace of technological advancement and the volatile nature of digital media, electronic records of enduring value will be lost if they are not identified for preservation and managed accordingly. If this issue remains unresolved, organizations face a future digital dark age where valuable information assets and cultural history are lost forever. This paper explores the digital preservation need with respect to public higher education institutions through literature review and a survey of practitioners. Several major government-sponsored projects were reviewed, providing insights into the complexity and scale of the issue with none of the projects able to fully address all issues. Results indicate that organizations are insufficiently prepared to identify electronic records of enduring value so that digital preservation techniques may be applied. Recommendations are provided for Information Systems professionals to address this burgeoning need.

Keywords

Digital preservation, digital archival repository, technological obsolescence.

INTRODUCTION

Organizations succeed and fail based upon their ability to effectively collect and utilize information. An organization's most valuable information is generally referred to as a record and can consist of documents such as annual reports, company policies, invoices, and personnel files. Many records have long-term historic value and this is especially true of government records. Although paper records can last for hundreds of years and still be readable, there is mounting concern that electronic records of enduring value will be lost due to the rapid pace of technological advancement and the volatile nature of digital media. Magnetically stored bits and dye-based DVDs degrade over the years, rendering the 1's and 0's unreadable. There is also concern that, despite the availability of digital preservation techniques (Lee, Slattery, Lu, Tang and McCrary, 2002), records will be lost to technological obsolescence as hardware and software is lost that can read the media, decode the data, and reproduce it in sufficient fidelity.

PROBLEM STATEMENT AND RESEARCH QUESTIONS

Many organizations struggle with the management of their electronic records. One issue which remains a significant challenge is the need to ensure future accessibility of electronic records with enduring historical value in the face of a recurring threat of technological obsolescence. Mitigation techniques exist but records creators, archivists, and Information Systems professionals must collaborate to first identify the affected records and then apply the digital preservation techniques. Digital government records, in particular, are a source of concern. Many government records have rich historic value over centuries and even millennia. Electronic records of enduring value are being generated at an accelerating pace, leaving unprepared organizations vulnerable. If this issue remains unresolved, organizations face a future digital dark age where valuable information assets and cultural history are lost forever.

Most records produced by an organization are of short or long-term value but not of enduring or permanent value. However, oftentimes the records of enduring value are stored in mixed environments. Filing cabinets or electronic file shares must be reviewed folder-by-folder and, sometimes, file-by-file to identify and retain. Typically, fewer than five percent of records are of enduring value (Keller, 2009) with an office disposing of about 350 pounds of paper per employee each year (Natural Resources Defense Council, n.d.). This can make identification of records of enduring value difficult in paper or electronic format, leading to over-retention of low-value records or accidental destruction of important historical records. In addition, records producers oftentimes are unaware of what may be of enduring value, particularly as such valuation is subjective. By the same token, archivists oftentimes do not possess sufficient context with respect to records in order to make a fully informed appraisal of the records' value. To help address these issues, I have organized my research around two questions:

- Given the need to apply digital preservation techniques to electronic records of enduring value to mitigate technology volatility, how can Information Systems professionals at public higher education institutions help records creators and archivists identify such records created and stored in diverse operational systems?

- Can Information Systems professionals across distributed units of a large institution collaborate to provide a cohesive approach to the preservation of electronic records of enduring value?

LITERATURE REVIEW

Analog files such as paper files or photographs require only the ability to see and the ability to read the language in order to understand the information it contains. With digital information, however, both hardware and software are also needed to translate the data bits into human-readable information (Bollacker, 2010). Many stone and paper records have lasted for thousands of years while electronic media typically degrades within decades and, because so much information is stored in a small space, media degradation can affect a large amount of information (Bollacker, 2010). To enable long-term preservation, it is recommended to use a widely adopted software and media technologies, use software emulators to run old programs on newer hardware, regularly migrate files to new media and new formats, store copies of files in geographically dispersed locations, and possibly even store a copy of files in a non-digital format such as paper for very important information to enable long-term preservation (Bollacker, 2010).

Issue 1: Can Technology Solve the Problem?

Many products are marketed with the claim that, by deploying them, they will allow an organization to easily identify and capture the information as well as support it indefinitely. This belief has even entered academia with one publication describing a particular technology as "An End to the Digital Dark Age" (Panos, 2003). Although auto-capture technologies serve a useful purpose, they are not capable of supporting an individual organization's need to capture an intelligent subset of high-value content based upon subjective measures of historic value. Such auto-capture tools typically can't capture isolated, password-protected, intranet, dynamic, or non-web content (Panos, 2003).

The National Archives and Records Administration (NARA) has even faced challenges in its attempts to develop an Electronic Records Archive (ERA) due to scope issues and the attempt to have one solution to meet all capture and preservation needs. According to a United States Government Accountability Office (GAO) report from 2009, NARA has spent $237 million over a seven year period to complete two of five increments of the ERA (Powner, 2009). The project has faced cost and time overruns and is expected to cost $550 million when complete. However, the GAO report indicates significant risks to the project while functionality deferments to later phases are cause for continued concern.

Four large European Union-funded digital preservation projects, ARCOMEM, SCAPE, ENSURE, and TIMBUS, address different aspects of the capture and preservation challenges including what data is being preserved, how it is identified, and how it is preserved (Edelstein, Factor, King, Risse, Salant and Taylor, 2011). According to Edelstein et al., the ARCOMEM project attempts to increase the social media and Web 2.0 capabilities of archives and archival processing. SCAPE addresses scalability of digital archives through development of tools, workflow automation, and integrated planning and control system for digital archives management. ENSURE supports healthcare with preservation options based on an organization's preservation needs, capabilities, and budget. Finally, TIMBUS attempts to address the preservation needs of digital workflows, particularly in the face of outsourced processes such as Software as a Service (SaaS) (Edelstein et al., 2011). For example, the process by which a highway bridge maximum load is calculated may be desirable to preserve to enable future load calculations or as part of a forensic analysis of a structural failure.

All four projects recognize human intervention as a requirement for determining what to preserve as well as for planning and execution of system upgrades or replacements (Edelstein et al., 2011). Scalability, especially in process automation, is a concern addressed by most of the projects due to volume. Most of the projects also attempt to automate quality control of the preservation lifecycle. SCAPE uses web-identified trends in planning for automated preservation processes. The use of cost and risk management approaches to digital archival processing is somewhat novel for a traditional archive.

A related project described in the digital libraries industry publication, D-Lib Magazine, noted the technical requirements for the various stages of preservation including preservation of the object itself, the context, the scalability, and the data sources (Moore, Baru, Rajasekar, Ludaescher, Marciano, Wan, Schroeder and Gupta, 2000). When applied to a specific collection, Moore et al. (2000) provided a proof of concept for such a system but thirty technical issues remained. In another D-Lib article, a digital repository at Stanford University is described including the lessons learned during its first five years (Cramer and Kott, 2010). Stanford was able to operate in production for three years, ingesting several hundred thousand objects and migrating many of them to new media. Although this proved successful, numerous issues during this time highlighted the need for significant revisions to the model, architecture, and strategy (Cramer and Kott, 2010). Stanford's system excludes the archival capture and accessioning processes from its repository, requiring those activities to occur prior to deposit in the repository.

A framework to evaluate various applications for their preservation capabilities, the Audit Checklist for the Certification of a Trusted Digital Repository (Kaczmarek, Hswe, Eke and Habing, 2006), applies a traditional software evaluation methodology

to digital archival repositories by using the Audit Checklist as the basis for scoring to assess a particular system's compliance with the Trusted Digital Repository standard. The specific weighting method utilized for evaluations will vary based upon institutional goals and needs.

Issue 2: Can Behavioral Change Solve the Problem?

Some have suggested that a successful approach to the preservation of electronic records must involve a change in business and archival processes, workflows, and procedures. One intriguing procedural concept is a mechanism by which at-risk electronic records could be pre-identified and proactively and aggressively saved in the event of a catastrophe.

A Task Force on Archiving of Digital Information report states that, without the necessary organizational support, "preservation of the nation's cultural heritage in digital form will likely be overly dependent on marketplace forces, which may value information for too short a period and without applying broader, public interest criteria" (Garrett and Waters, 1996). The report goes on to say that "initial responsibility for preservation begins with creation of the information and rests with the creator, owner or provider of it" (Garrett and Waters, 1996). Under this concept, the initial electronic record custodian is the first archivist for the record and must provide essential retention, categorization, and migration services in the early years of preservation. However, this person is typically not trained as an archivist or in electronic preservation strategies. Therefore, education of electronic record custodians is needed to ensure that electronic records of enduring value will not be lost physically or functionally prior to archival accessioning.

A distinction can be made between working, presentation, and archival forms of digital files (Simons, 2004). According to Simons, most researchers focus on the working and presentation formats while not considering the archival format. In order to ensure that research data is accessible for very long periods of time, he encourages grant funders to require archival formats for research output, establish incentives for archival consideration in peer review and promotion and tenure systems, and for researchers to become more closely involved with libraries and archives with respect to digital preservation (Simons, 2004).

In an article in Learned Publishing, the authors state that, while over 90% of journals have a digital preservation policy, few journals have preservation strategies for journal accompaniments such as research datasets or multimedia (Smit, Van Der Hoeven and Giaretta, 2011). Citing an EU project report, the authors list seven mostly non-technology threats to electronic records and eight behavioral, non-technology based aspects of electronic research data preservation.

Issue 3: Do Content Creators Know What Is of Enduring Value?

A final major issue highlighted by the literature review was an underlying assumption that records creators and custodians had sufficient knowledge of their files to know what materials are of temporary value and what materials are of enduring value. In an effort to prune out low-value records, there is the risk that valuable records will inadvertently be destroyed as well due to poor organization, poor oversight of the disposal process, or poor understanding of the differences between archival and non-archival records. The confusion between stakeholders, even among archivists, is identified as a significant challenge for a New Zealand digital archive project (Knight, 2010).

A 2011 report for the Digital Preservation Coalition discusses the institutional attitudes of account size or chronological limitations as the most popular controls on email growth (Prom, 2011). Prom suggests several approaches to the preservation of email messages, depending on institutional and individual preferences, while supporting operational needs to manage email costs and risk. The first approach is preservation of an entire account or system without consideration of record or non-record content. The second approach is more collaborative with archivists working with account holders to better identify and separate record from non-record content. Another approach is a message-by-message approach for individual account owners to manage their content (Prom, 2011).

METHODOLOGY

To accomplish the goals of this research, I reviewed numerous resources including government publications, project reports, and scholarly journals on archival and information systems topics. In addition, experts from the Illinois State Archives and University of Illinois' Archives, Counsel, Security, and Information Technology offices were surveyed to complement the literature review (see Appendix A for the online survey questions). This approach provided significant insight in an informal setting both from internationally-recognized bodies as well as esteemed practitioners as key informants. Finally, instructor and peer feedback was an important resource to improve upon research quality.

I analyzed the data collected from the surveys through qualitative and quantitative methods, primarily through summarization and through coding concepts of particular note or that were also mentioned in the literature review. For the risk and preparedness rankings, I calculated the average across all respondents as well as compared that average to individual data points to determine the standard deviation. A high standard deviation might warrant further investigation depending on the department within the

organization that the individual represents. In addition, I compared the responses on qualitative questions based on the individual's position, for instance giving greater weight to an archivist's response to an archival question and greater weight to an IT manager's response to an IT question.

With the results of the surveys analyzed, I then compared the results with the findings from the literature review (see the result section below). A comparison between scholarly articles, industry resources, and practitioner opinions was performed in order to determine if consensus exists regarding the pressing issues and potential solutions for electronic archival records systems. Similarities and differences were noted and reflected upon to better understand their meaning. Finally, a recommendation for organizations was developed as well as topics for additional exploration by scholarly researchers.

RESULTS

The analysis of quantitative questions from the five surveys can be seen in Figure 1. Responses indicate close alignment with one another across various roles within public records management. They also indicate a perception of insufficient preparedness and elevated risk of loss which aligns with information collected from the literature review including capability deficiencies noted by NARA and the EU project report's list of the seven threats to electronic records.

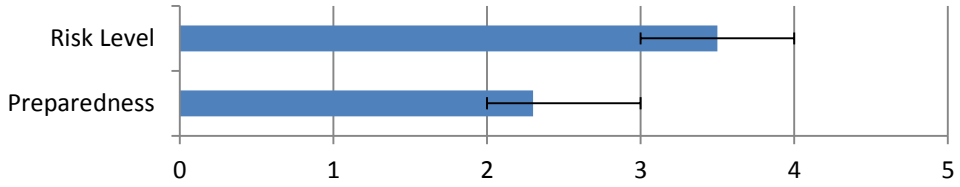

Figure 1. Risk Versus Preparedness

The first qualitative question asked respondents if their organization currently preserves electronic records. Non-archival staff responded with either specific IT systems that possess such records or pointed generally to archival operations. One archivist provided many different solutions currently being utilized as part of a piecemeal solution (Archives Representative, 2013) while another indicated that they possess but are not actively preserving electronic records (State Archives Representative, 2013).

The second qualitative question addresses the specific threats to the records. Tellingly, all respondents listed awareness by records custodians of the record's historical value and needs to preserve them was listed as one of the top issues. Risks due to technological obsolescence were also mentioned by multiple respondents. One respondent mentioned a few specific issues such as "…IT policies for automatic deletion, and poor indexing/organization of organically created e-records -The lack of a distinct plan thus far for the long-term management of the e-records in our care…" (State Archives Representative, 2013). The issue of automatic deletion policies was a particularly noteworthy threat, especially with email systems where most organizations have automated off boarding processes which delete entire email accounts of former employees after a defined period of time. This is a risk to content in those accounts which may have enduring value, particularly within accounts of senior administrators.

The third qualitative question addresses the resource needs to address these threats. Generally, these revolved around education, funding, and policy needs to provide the sufficient direction, administrative support, staff, infrastructure, and awareness over an extended timeframe that is necessary to provide a long-term solution.

In terms of identifying records of enduring value from their active systems, most respondents indicated that better understanding of existing policies and appraisals was necessary with some recommending that further refinement of definitions is necessary as well as assessments on a departmental basis based on job functions. One respondent mentioned "…functional analysis-type appraisal, where we identify potentially important records at the beginning, based upon business functions of a given unit" (State Archives Representative, 2013).

The impacts of the Illinois Freedom of Information Act (FOIA) mentioned in the surveys primarily addressed the need to efficiently store the records in order to quickly locate and review the records in order to respond to requests in the time required by the law. One respondent indicated that the law may discourage preservation (IT Security Representative, 2013) which is a threat mentioned by another respondent as those in possession of records seek to limit their liability and exposure to public release of information that reflects poorly on the institution (State Archives Representative, 2013).

Two respondents indicated that it may be possible for their organization to meet the needs of all archival electronics records and patrons using a single repository (IT Manager Representative, 2013 and Counsel Representative, 2013). Management of the repository including access controls was important, though. The three archivists all indicated that it was not likely possible

to utilize a single repository. One archivist provided reasons such as difficulty to handle irregular high demand and the need to address very specific record needs such as Computer Automated Drafting (CAD) files (State Archives Representative, 2013).

The respondents' perceived impacts of public access to electronic records varied quite a bit from increased workload for archives staff to improved public research to self-censoring by public employees. One respondent indicated that the length of time elapsed between the creation of a record and its release for public access is critical to ensure that there isn't a chilling effect on what gets recorded (Counsel Representative, 2013). This may be mitigated by lengthy delays in access, mediating researcher access to the materials through a trusted gatekeeper such as an archivist or public access coordinator, pre-filtering the material based upon various software tools to redact sensitive information, or limiting access to the materials to offline.

The final question asked respondents if their organization would consider outsourcing the preservation of electronic records to another institution. One archivist did not feel that outsourcing would be something her organization would be willing to do while another indicated that certain functions such as storage of redundant copies could be outsourced such as to a cloud provider if cost effective (State Archives Representative, 2013). Although outsourcing may be an option for many, it appears to be too much in its infancy as an option for most archival operations to rely upon too heavily. This may be due to the conservative nature of the archival profession, past experience with technology failure, unfamiliarity with new technology, or a combination of factors.

SUMMARY AND CONCLUSIONS

In reviewing the results of the literature review, surveys, and analysis, the following conclusions were made:

1. Technology can enable preservation of electronic records but it is not a turnkey solution. Many automated tools exist which can help to address issues such as redaction, search and e-discovery, indexing, and even format migration. In addition, the tools are not all part of a single, integrated suite so interoperability factors must be considered. Information Systems professionals should coordinate to promulgate enterprise standards and tools to support digital preservation and access needs.
2. Archivists must align themselves with the existing processes and structure in order to minimize disruption to business operations. In doing so, they will be more likely to be successful in obtaining voluntary participation in the identification, capture, processing, and preservation of electronic records of enduring value. Attempts to change business processes to better enable archival needs will likely be met with significant resistance as unnecessary and costly intrusion into core business functions.
3. The need for education and training to help content creators better understand, identify, and categorize their content is essential. Most users seek help managing their records, both of enduring and temporary value. Traditionally, this help has been sought out as space available for paper records became limited. With digital files, however, help is more often sought out due to concerns over retrieval speed, risk mitigation, or simply to clean-up old files, often prompted by staff turnover. This provides an excellent opportunity for Information Systems professionals to help offices manage their information assets, increase efficiency and productivity, and create a digital preservation plan their records of enduring value.

RECOMMENDATION FOR FUTURE RESEARCH

Although much insight was gained from this paper, it also highlighted additional areas that warrant further exploration. There is a need for evaluation of specific digital archival repository technologies to understand the current capabilities, limitations, and needs of institutions seeking to preserve their electronic records of enduring value. In addition, there is a need for well-defined approaches for the identification and capture of electronic records of enduring value with a low operational impact to ensure that such records are not lost. Finally, additional research is needed to enumerate and address the technological and data security risk factors affecting long-term digital preservation including how information systems professionals can help mitigate these risks.

REFERENCES

1. Archives Representative. (2013) Survey by Brent West. University of Illinois.
2. Bollacker, K. (2010) Avoiding a digital dark age, *American Scientist,* 106-110.
3. Counsel Representative. (2013) Survey by Brent West. University of Illinois.
4. Cramer, T. and Kott, K. (2010) Designing and implementing second generation digital preservation services: A scalable model for the Stanford digital repository, *D-Lib Magazine*, 16, 9.

5. Edelstein, O., Factor, M., King, R., Risse, T., Salant, E. and Taylor, P. (2011) Evolving domains, problems and solutions for long term digital preservation, *Proceedings of the 8th International Conference on Preservation of Digital Objects*, November 1-4, Singapore.
6. Garrett, J. and Waters, D. (1996) Preserving Digital Information, *Task Force on Archiving of Digital Information*.
7. IT Manager Representative. (2013) Survey by Brent West. University of Illinois.
8. IT Security Representative. (2013) Survey by Brent West. University of Illinois.
9. Kaczmarek, J., Hswe, P., Eke, J. and Habing, T. (2006) Using the Audit Checklist for the Certification of a Trusted Digital Repository as a Framework for Evaluating Repository Software Applications, *D-Lib Magazine*, 12, 12.
10. Keller, C. (2009) Records Retention and Disposition, *EDUCAUSE Wiki*, http://www.educause.edu/wiki/records-retention-and-disposition (accessed November 19, 2013).
11. Knight, S. (2010) Early learnings from the National Library of New Zealand's National Digital Heritage Archive project, *Program: electronic library and information systems*, 44, 2, 85-97.
12. Lee, K., Slattery, O., Lu, R., Tang, X. and McCrary, V. (2002) The State of the Art and Practice in Digital Preservation, *Journal of Research of the National Institute of Standards and Technology*, 107, 1, 93-106.
13. Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W. and Gupta, A. (2000) Collection-Based Persistent Digital Archives, *D-Lib Magazine*, 6, 3.
14. Natural Resources Defense Council. (n.d.) Reducing Paper Use, *Greening Advisor*, http://www.nrdc.org/enterprise/greeningadvisor/pa-reducing.asp (accessed November 19, 2013).
15. Panos, P. (2003) The Internet Archive: An End to the Digital Dark Age, *Journal of Social Work Education*, 343-347.
16. Powner, D. (2009) NATIONAL ARCHIVES: Progress and Risks in Implementing its Electronic Records Archive Initiative, *U.S. House of Representatives Testimony*, Washington, D.C.
17. Prom, C. (2011) Preserving Email, *Technology Watch Report*, Digital Preservation Coalition.
18. Simons, G. (2004) Ensuring that digital data last: The priority of archival form over working form and presentation form, *E-MELD Symposium on "Endangered Data vs. Enduring Practice"*, Boston, Linguistic Society of America.
19. Smit, E., Van Der Hoeven, J. and Giaretta, D. (2011) Avoiding a Digital Dark Age for data, *Learned Publishing*, 35-49.
20. State Archives Representative. (2013) Survey by Brent West. Illinois State Archives.

APPENDIX A: SURVEY QUESTIONS

1. Does your organization, university, or State agency currently preserve electronic records? If so, how?
2. How prepared is your organization to address the long-term preservation of electronic records? (5 = Very Prepared)
3. What current and future threats exist for electronic records of enduring value at your organization?
4. How at risk are your organization's electronic records of enduring value to loss? (5 = High Risk)
5. What resources would enable your organization to better mitigate these threats?
6. How can electronic records of enduring value be identified from their active system?
7. What laws impact your organization's ability to create and maintain electronic records?
8. How does the Freedom of Information Act impact electronic records of enduring value?
9. Can a single repository serve the needs of all archival electronic records and all patrons?
10. What operational or cultural impacts does/would the availability of electronic records to archives patrons produce on employees or the public?
11. Would your organization consider outsourcing preservation management to another public or private institution?